UNIVERSITY OF BIRMINGHAM
University of Birmingham Research at Birmingham

Data driven region-of-interest selection without inflating type I error rate

Brooks, Joseph; Zoumpoulaki, Alexia; Bowman, Howard

DOI:
10.1111/psyp.12682

License:
Creative Commons: Attribution-NonCommercial (CC BY-NC)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Brooks, J, Zoumpoulaki, A & Bowman, H 2017, 'Data driven region-of-interest selection without inflating type I error rate', *Psychophysiology*, vol. 54, no. 1, pp. 100-113. https://doi.org/10.1111/psyp.12682

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

•Users may freely distribute the URL that is used to identify this publication.
•Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
•User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
•Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 04. May. 2023

PSYCHOPHYSIOLOGY

Psychophysiology, 54 (2016), 100–113. Wiley Periodicals, Inc. Printed in the USA.
Copyright © 2016 The Authors. Psychophysiology published by Wiley Periodicals, Inc. on behalf of Society for Psychophysiological Research
DOI: 10.1111/psyp.12682

Data-driven region-of-interest selection without inflating Type I error rate

JOSEPH L. BROOKS,[a,b] ALEXIA ZOUMPOULAKI,[b,c] AND HOWARD BOWMAN[b,c,d]

[a]School of Psychology, University of Kent, Canterbury, UK
[b]Centre for Cognitive Neuroscience and Cognitive Systems, University of Kent, Canterbury, UK
[c]School of Computing, University of Kent, Canterbury, UK
[d]School of Psychology, University of Birmingham, Birmingham, UK

Abstract

In ERP and other large multidimensional neuroscience data sets, researchers often select regions of interest (ROIs) for analysis. The method of ROI selection can critically affect the conclusions of a study by causing the researcher to miss effects in the data or to detect spurious effects. In practice, to avoid inflating Type I error rate (i.e., false positives), ROIs are often based on a priori hypotheses or independent information. However, this can be insensitive to experiment-specific variations in effect location (e.g., latency shifts) reducing power to detect effects. Data-driven ROI selection, in contrast, is nonindependent and uses the data under analysis to determine ROI positions. Therefore, it has potential to select ROIs based on experiment-specific information and increase power for detecting effects. However, data-driven methods have been criticized because they can substantially inflate Type I error rate. Here, we demonstrate, using simulations of simple ERP experiments, that data-driven ROI selection can indeed be more powerful than a priori hypotheses or independent information. Furthermore, we show that data-driven ROI selection using the aggregate grand average from trials (AGAT), despite being based on the data at hand, can be safely used for ROI selection under many circumstances. However, when there is a noise difference between conditions, using the AGAT can inflate Type I error and should be avoided. We identify critical assumptions for use of the AGAT and provide a basis for researchers to use, and reviewers to assess, data-driven methods of ROI localization in ERP and other studies.

Descriptors: ERPs, EEG, Analysis/statistical methods

Analysis of neuroimaging data (e.g., EEG, magnetoencephalography [MEG], MRI) can involve hundreds or thousands of statistical tests. A significant challenge in analysis of such data is how, with high power, to detect effects without increasing the Type I error (false positive) rate. Given that experiments typically show effects only for a small subset of the recorded data, one common approach is to select a region of interest (ROI) across one or more dimensions in the data. Correct identification of the ROI is often critical to the results of the study. If it is chosen incorrectly, then relevant effects may be missed, inflating the Type II error rate. On the other hand, if many locations are tested simultaneously (mass univariate) without proper correction or biased procedures are used for ROI selection (Kilner, 2013; Kriegeskorte, Simmons, Bellgowan, & Baker, 2009), then this can inflate the Type I error rate (i.e., false positives). Inflation of Type I error rate, along with low power (Button et al., 2013) and publication bias (Easterbrook, Gopalan, Berlin, & Matthews, 1991; Rosenthal, 1979), are serious issues that have significant knock-on consequences for the reliability of the scientific literature (Colquhoun, 2014).

ROIs are commonly selected using a priori hypotheses or based on independent data (Kilner, 2013; Luck, 2014). For instance, boundaries of an ROI for an ERP study of the face-sensitive N170 component (e.g., 150–190 ms., electrodes P7/P8) may be based on the ROI used in or location of significant effects in a previous study (e.g., Towler & Eimer, 2014). This approach makes no reference to features of the data under analysis, and it is safe and unbiased (i.e., does not inflate Type I errors) because ROI selection cannot be driven by noise fluctuations in the data (Kilner, 2013; Luck, 2014). This approach is widely used in ERP and event-related field (ERF in MEG) research.

However, there can be significant variation in the temporal or spatial location of effects between experiments due to differences

Thanks to members of the cognitive group meeting at University of Kent for critical feedback and to Jason Arita at UC Davis for advice on ERPLAB functions.

[The copyright line for this article was changed on 20 January 2017 after original online publication.]

Address correspondence to: Joseph L. Brooks, School of Psychology, Keynes College, University of Kent, Canterbury, CT2 7NP, United Kingdom. E-mail: J.L.Brooks@kent.ac.uk

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

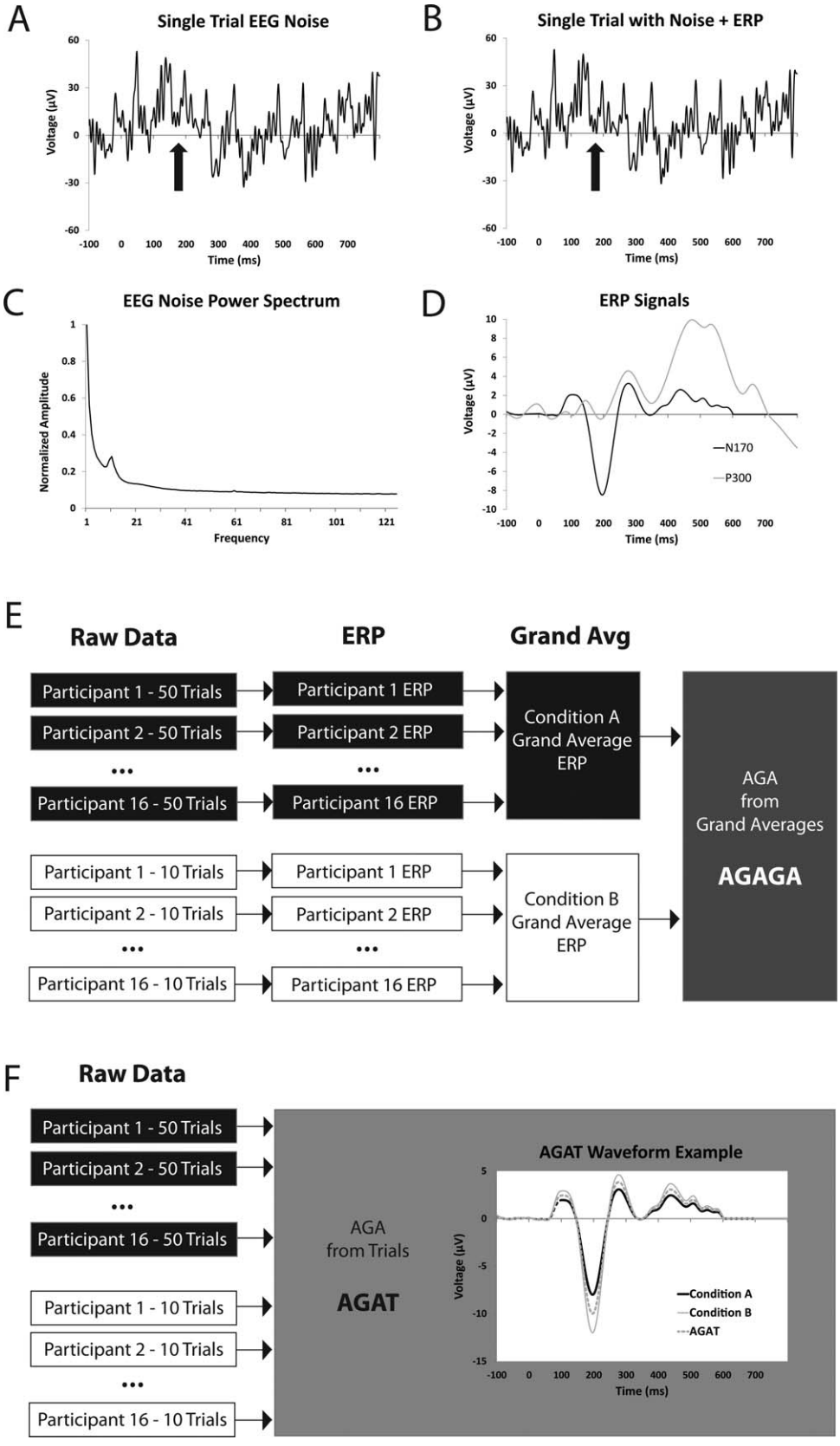

Figure 1. Examples of simulated data and calculation of AGAT and AGAGA waveforms. A: Example of an EEG pure noise waveform for an individual trial. B: Some simulations contained both noise and ERP deflections. The arrow below each waveform indicates a point of difference between (A) and (B) caused by the addition of the N170 ERP to the signal in (B). C: Power spectrum of EEG data used to scale the amplitudes of sinusoids in the creation of EEG noise. D: Pure ERP signal waveforms (without noise) for the N170 (black) and P300 (gray), which were added to single trials in

Continued.

in design, stimulus characteristics (e.g., Flevaris, Robertson, & Bentin, 2008; Zhang & Luck, 2009), and unknown noise factors. For example, the attention-related ERP component, N2pc (Luck & Hillyard, 1994), appears later in time for weaker stimuli than for stronger stimuli (e.g., Brisson, Robitaille, & Jolicoeur, 2007). Although precedents for such stimulus-based effect shifts may be available in some cases, this will often not be the case, especially because the point of many experiments is to study an ERP component under novel conditions. Furthermore, even when precedents are available, there can be several different options (especially for well-studied effects), often with no clear rationale for choosing among them. This provides opportunities for post hoc "fishing" and, without correction, can inflate Type I error rates (Simmons, Nelson, & Simonsohn, 2011). ROI selection based on hypotheses or independent data cannot usually account for interexperiment variation, and this may reduce the probability of detecting an effect.

For optimal detection of effects, the ROI selection process should be sensitive to experiment-specific features of the data, that is, data driven. A data-driven approach would use features of the data under analysis to position the ROI. In the N170 example above, data-driven ROI selection may, for instance, search through the observed data in space and time and position the ROI at the largest negative peak within a predetermined time period (e.g., Caharel et al., 2013), for example, 120–240 ms (de Gelder & Stekelenburg, 2005), and spatial window on the scalp. This would allow the ROI selection process to account for the experiment-specific location of the N170-associated peak. This may or may not overlap with the locations of previous findings. Although peaks are common and easily quantifiable features of interest in ERP studies, this is by no means the only relevant, or even appropriate, feature for data-driven analysis (Luck, 2005, 2014). Other more sophisticated features have been used (Koenig, Stein, Grieder, & Kottlow, 2014; Ten Caat, Lorist, Bezdan, Roerdink, & Maurits, 2008). The appropriate feature should be determined by hypothesis, theory, or a priori assumptions. We focus on peaks here because they are commonly used and easily quantifiable.

Data-driven approaches to ROI localization, especially but not only in ERP research, have faced criticism that they can inflate Type I error rates (Kilner, 2013; Kriegeskorte et al., 2009; Luck, 2014; Vul, Harris, Winkielman, & Pashler, 2009). Publication guidelines (Keil et al., 2014) and methods books (Luck, 2014) specifically warn about the dangers of this type of ROI localization. This is because the data features used for selection (e.g., a peak) can be affected by random noise. If this noise is not independent of the contrast of interest (e.g., difference between conditions), then using it for ROI selection will inflate Type I errors. Similar issues have arisen and garnered significant attention in fMRI (e.g., Kriegeskorte et al., 2009; Vul et al., 2009) and exploratory behavioral research (e.g., Simmons et al., 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Nonetheless, we believe that some researchers already employ some form of data-driven approach despite the fact that there are few, if any, published and empirically validated data-driven ROI selection procedures for ERP data. For instance, some researchers select peaks on what we will call the aggregate grand average of grand averages (AGAGA). In a simple experiment with two conditions, this is simply the average of the two condition grand-averaged waveforms (Figure 1E). However, whether, and under which conditions, this waveform is unbiased is not completely clear. This leaves room for incorrect use, which will inflate false positive rates. Thus, to avoid criticism, many researchers may avoid using data-driven methods altogether. This has the consequence of missing opportunities to increase power.

Our goal is to demonstrate empirically that data-driven ROI selection can be used safely in ERP (and, by extension, other) experiments and thereby take advantage of study-specific information to reduce Type II errors, while still maintaining Type I error rate at 5%. We will focus on ERP data because ROIs are routinely used in ERP analysis, ERP work forms a large body of cognitive neuroscience research, and because recent criticism suggests that data-driven approaches used in this area may be biased or at least poorly reported (Kilner, 2013, 2014). However, the basic issues apply in principle to other types of data in which ROIs are used, and similar issues can arise, for example, MEG ERFs, psychophysiology, eye tracking (e.g., von der Malsburg & Angele, 2015).

To perform ROI selection, we will compute what we call the aggregate grand average from trials (AGAT), which is similar to the use of orthogonal contrasts for ROI selection in fMRI research (Kriegeskorte et al., 2009), and demonstrate that selection of ROIs based on this waveform is unbiased and does not inflate Type I error rates. In the simplest case, the AGAT is computed by aggregating all of the individual trial waveforms/time series from all participants and conditions and averaging across them to form a single time series (Figure 1F), the AGAT. It is important to notice that the AGAT is, in some circumstances (see Simulation 2), distinct from the AGAGA, described above, which is more naturally derived from the typical ERP processing pipeline (Figure 1E). In this study, we will show that AGAT-based ROI selection is safe for both balanced (Simulation 1) and unbalanced designs (i.e., different amounts of data between conditions, Simulation 2), demonstrate conditions under which it can fail (Simulation 3), and establish its power relative to widely used ROI selection based on independent data (Simulation 4). Importantly, we will also examine some of the assumptions that are critical for proper use of the AGAT method and which are also likely relevant to other ROI selection methods. In particular, use of the AGAT may not be effective if the waveform morphology or latency of ERP features of interest (e.g., peaks) differ substantially between the conditions (see Discussion for more detail). The results and interpretation of these simulations will empower researchers and reviewers to make educated decisions about data-driven ROI selection and, hopefully, prompt further discussion and method development in this domain. To

simulations containing ERP deflections. Note the different scale from (A) and (B). E: The aggregate grand average from grand averages (AGAGA) was computed by averaging the individual trials separately within each condition (Condition A in black boxes, Condition B in white boxes) for each participant into an ERP waveform for each participant. Then, these participant ERPs were averaged within each condition to form a grand-averaged ERP for each condition. The AGAGA waveform was created by averaging the condition grand-averaged ERPs. Arrows indicate an averaging process. Note that (E) represents an experiment with a condition trial number asymmetry as in Simulation 2. However, most experiments will have approximately the same number of trials in each condition. F: The aggregate grand average from trials (AGAT) was created by aggregating all of the individual trials, from all participants and both conditions, into one group and then averaging them. An example of the AGAT waveform (dashed gray line) is plotted along with grand averages for the two conditions (thick black line and thin gray line). Note that the amplitude difference between conditions here is for illustration purposes only and was not present in null hypothesis simulations (Simulations 1–3).

support our claims, we will conduct null hypothesis data simulations, under various conditions, to assess the Type I error rate and also power simulations associated with using the AGAT for ROI selection in an ERP experiment with realistic EEG noise and two ERP deflections of different polarity (to show generalizability).

Simulation 1: AGAT Type I Error Rate

Simulation 1 focused on estimating the Type I error rate associated with using data-driven ROIs selected using the data-driven AGAT waveform. It compared this to other data-driven ROIs including the already discredited difference wave (Kilner, 2013) and the AGAGA. To test generality across different types of data, Simulation 1 used three different ERP signal types. One contained noise-only data (Simulation 1A). The other two had realistic ERP deflections (P300, Simulation 1B; and N170, Simulation 1C) added to the noise in individual trials so that the grand averages contained ERP-like waveform morphology. It is not practically possible to simulate all possible ERP waveform types. However, by using these three different types of ERP data (noise-only, negative polarity ERP, and positive polarity ERP), including two widely used ERP components, we aimed to test whether our conclusions about the safety of the AGAT are significantly affected by the exact morphology and polarity of the ERP waveform. We expected that the AGAT-based ROIs will maintain Type I error rates at 5%, whereas selecting ROIs based on the difference wave will substantially inflate Type I error rates.

Method

We performed 12 versions of Simulation 1 in R (R Development Core Team, 2014), version 3.1.0. These 12 versions arose from varying two orthogonal factors. First, we varied the signal content of the data: (Simulation 1A) EEG noise-only, (Simulation 1B) noise+P300, and (Simulation 1C) noise+N170. Within each of these three versions, we also created four variations with different numbers of channels in the data (1, 8, 16, or 32). The label Simulation 1A refers to the class of all simulations containing noise-only data. The label 1A-16Ch refers to the single simulation involving noise-only data with 16 channels. For each individual simulation, we generated data for 10,000 experiments, each having two conditions with 16 participants, 50 trials per condition, and either 1, 8, 16, or 32 channels of data. Each trial comprised 900 sample points with a sampling rate of 1000 Hz and time points −100 to 800 ms. The EEG noise time series (e.g., Figure 1A) for each individual trial was generated by summing 50 sinusoids with randomly (without replacement) chosen frequencies (integer values 1–125 Hz) and random phases (with replacement, different across frequencies and trials), 0–2 π (Yeung, Bogacz, Holroyd, & Cohen, 2004). Each sinusoid was scaled according to its frequency's power in the human EEG power spectrum (Figure 1C) and normalized to the 1 Hz amplitude. The resulting noise waveform was multiplied by 20 μV to increase its overall amplitude. The noise in each channel was created independently without spatial or temporal autocorrelation.

For Simulations 1B and 1C with ERP signals, we added one of the ERP signal waveforms (Figure 1D) to the EEG noise (produced as above) on each trial (e.g., Figure 1B), equivalently in both conditions. ERP waveforms were derived from grand averages in previous studies in our group: P300 (fake condition in Bowman et al., 2013) and N170 (unpublished data).[1] The ERP peak amplitudes were scaled such that the maximum for P300 was at 8 and the minimum for N170 was at −8. This was done to ensure that signal-to-noise ratio of the two signals was equivalent.

For each of the 10,000 experiments within a simulation, we derived three waveforms to be used in ROI selection: the difference wave, the AGAGA, and the AGAT. The difference wave was calculated by, within each Condition A and B, creating participant ERPs (i.e., averaging across trials within each condition for each participant, see Figure 1E) and then averaging these participant ERPs into a grand average for each condition, GA_A and GA_B. The difference wave was the subtraction of the two grand-averaged waves, $GA_A - GA_B$. The AGAGA was calculated by averaging the two grand-averaged waveforms. The AGAT waveform was calculated by aggregating all of the individual trial waveforms from all participants and both conditions into a single group (i.e., 2 Conditions × 50 Trials × 16 Participants = 1,600 trials) and averaging the waveforms (Figure 1F). In Simulation 1, the AGAGA and AGAT were equivalent.

The ROIs on the difference, AGAGA, and AGAT waves in each experiment were positioned for detecting the relevant peak: Simulation 1A (noise-only), minimum value (arbitrarily chosen); Simulation 1B (Noise+P300), maximum value to detect P300 peak; Simulation 1C (Noise+N170), minimum value to detect N170 peak. For data with ERP signals, the rule was chosen to identify the feature of interest in the AGAGA/AGAT (e.g., N170 peak is a minimum). For noise-only data, the rule was arbitrarily set to locate a minimum. Results were equivalent when we used a maximum rule for noise-only data. An unsigned rule was also implemented for noise-only data and produced equivalent results for the AGAT but further inflated the Type I error rate for difference wave-based ROIs. For data with more than one channel, the ROI was selected as the maximum or minimum across the two-dimensional Time × Channel Space and the ROI was centered at a channel-time coordinate.

We conducted an unpaired-samples t test between conditions at each ROI location. This used individual participants' ERP amplitudes at the ROI location (e.g., two groups of 16 amplitudes). We also conducted these t tests using four integration windows of different sizes (10, 20, 50, 100 samples) to understand their effect and to account for common practice of averaging over intervals/windows around an ROI center point. In these tests, voltage in each participant's ERP was averaged (across time) within the window centered at the ROI position. For each simulation, we estimated the Type I error rate for each combination of ROI type and integration window as the percentage of experiments with a significant difference between conditions. We computed 95% CIs of the Type I error rate in each simulation with the bootstrapping function in R using 5,000 bootstrap replicates and the "basic" bootstrap method. This involved resampling the original distribution of 10,000 p values and recalculating the Type I error rate for each replicate.

1. The data for the N170 waveform was derived from an experiment in which an ambiguous Rubin faces—vase stimulus was shown for 150 ms on each trial (followed by a white noise mask for 100 ms), and participants responded about whether they saw the face regions as figure or the vase region as figural. The N170 waveform was from data collapsed over the two response options and averaged across electrodes P10, P8, PO8, P9, P7, and PO7. There were 17 participants and 300 trials per participant. Data were recorded with a BioSemi ActiveTwo active electrode system and sampled at 1024 Hz with an average reference of the 64 scalp channels.

Results and Discussion

In Simulation 1, we estimated the Type I error rate associated with AGAT-based ROIs. As expected from previous work (Kilner, 2013), the Type I error rate for difference wave-based ROIs in all simulations consistently exceeded the desired 5% level with approximately 75% errors when using the smallest integration window (1 sample width). Type I error rate decreased as the integration window size increased (Figure 2A, light gray bars). In contrast, AGAT-based ROIs were associated with an approximate 5% error rate regardless of the integration window size and regardless of whether the data were pure noise (Figure 2A, dark gray bars) or contained ERP deflections (Figure 2B,C, dark gray bars). The AGAGA produced identical results to the AGAT and thus is not plotted separately. Figures 2A–C show results from simulations of one channel data. The AGAT results for multichannel data were equivalent. Figure 2D shows that, as the number of channels increased in Simulation 1C (N170 data), using an AGAT ROI maintained Type I error rate at 5%. In contrast, Type I error rate for the difference wave reached 100% as the number of channels increased. The multiple channel results were equivalent for Simulations 1A and 1B. We also conducted simulations where we varied the number of samples across time (i.e., increased sampling rate but with same length of time) and found that the AGAT maintained Type I error rate whereas the difference wave did not. Overall, our results suggest that the AGAT is safe regardless of the size of the data (number of Channels × number of Samples).

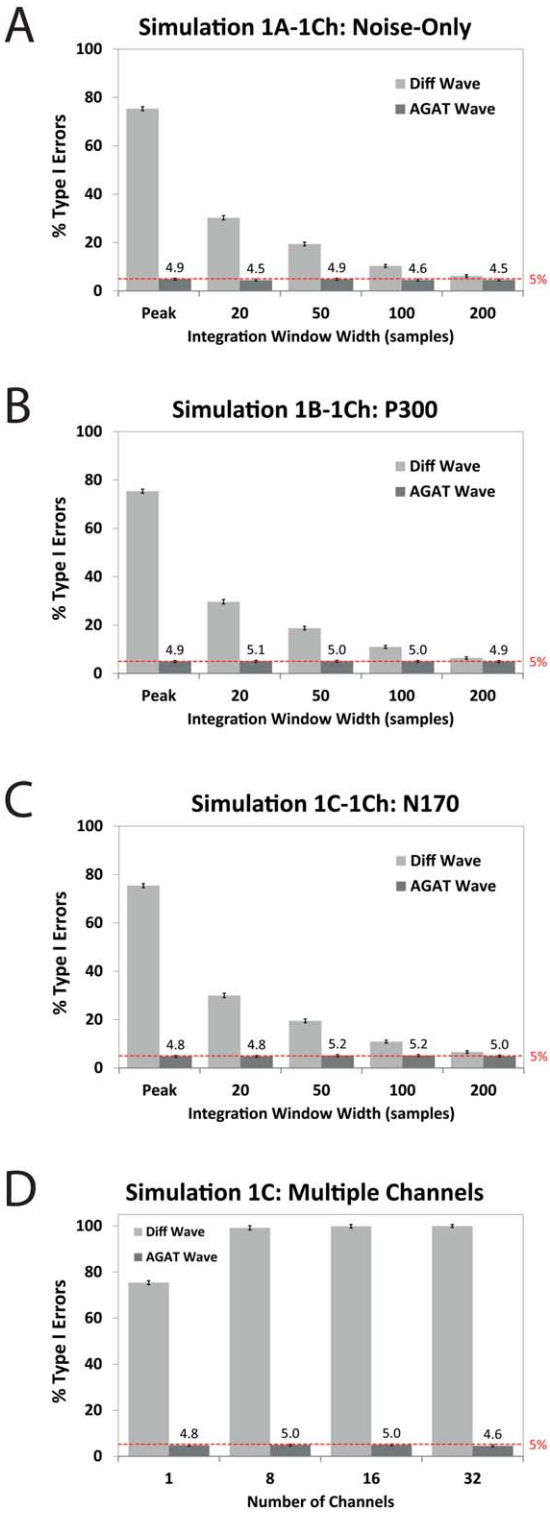

Figure 2. Simulation 1 results. The percentage of Type I errors is plotted as a function of the size of the integration window (in sample points) used for difference wave-based ROIs (light gray bars) and AGAT-based ROIs (dark gray bars). The horizontal dashed red line indicates the target 5% Type I error rate level. Error bars represent 95% CIs (see Method). A: Results for Simulation 1A-1Ch (single channel), noise-only show that AGAT-based ROIs maintain Type I error rate at 5% whereas the difference wave does not. B: Results for 1B-1Ch, Noise+P300 ERP. C: Results for 1C-1Ch, Noise+N170 ERP. Numbers above the AGAT-based dark bars indicate the percentage of Type I errors for those ROIs. D: Type I error rate results are plotted as a function of the number of channels in Simulation 1C (N170 data) for AGAT and difference wave ROIs. Results were similar for P300 and noise-only data (not shown here). Note that the maximum of the scale in (D) goes to 105% (100% in other panels).

These results clearly demonstrate that using data-driven AGAT-based ROIs does not inflate Type I error rate above 5%. This is because the AGAT time series is independent of the contrast of interest (i.e., the difference between conditions here). The average cross-correlation (zero-lag) coefficient between the AGAT and difference wave was not different from zero (one-sample t test) for any of the simulations: $r_{\text{Noise-Only}} = .003$, $t(9999) = 1.66$, $p = .09$; $r_{\text{P300}} = .0003$, $t(9999) = -0.16$, $p = .87$; $r_{\text{N170}} = .001$, $t(9999) = 0.60$, $p = .54$. Thus, the AGAT provides an unbiased, data-driven basis for ROI selection in ERP studies.

Simulation 2: Condition Trial Number Asymmetry

Unbiased performance of the AGAT ROI-selection procedure depends critically on it being independent of condition differences. Independence could be violated if the ROI-selection waveform were generated with unequal contributions of data from the two conditions, for example, mismatch negativity ERP component (e.g., Näätänen, Gaillard, & Mäntysalo, 1978). In this situation, the noise from one condition may be weighted more heavily in the AGAT than noise from the other condition, rendering the waveform nonindependent of condition differences. Using the same parameters as Simulations 1A–C, except now in the presence of a trial number asymmetry between conditions (and with only 1,500 experiments per simulation for computational efficiency), we estimated Type I error rates for ROIs based on the AGAT, the AGAGA, and the difference wave to test their performance under condition trial number asymmetry.

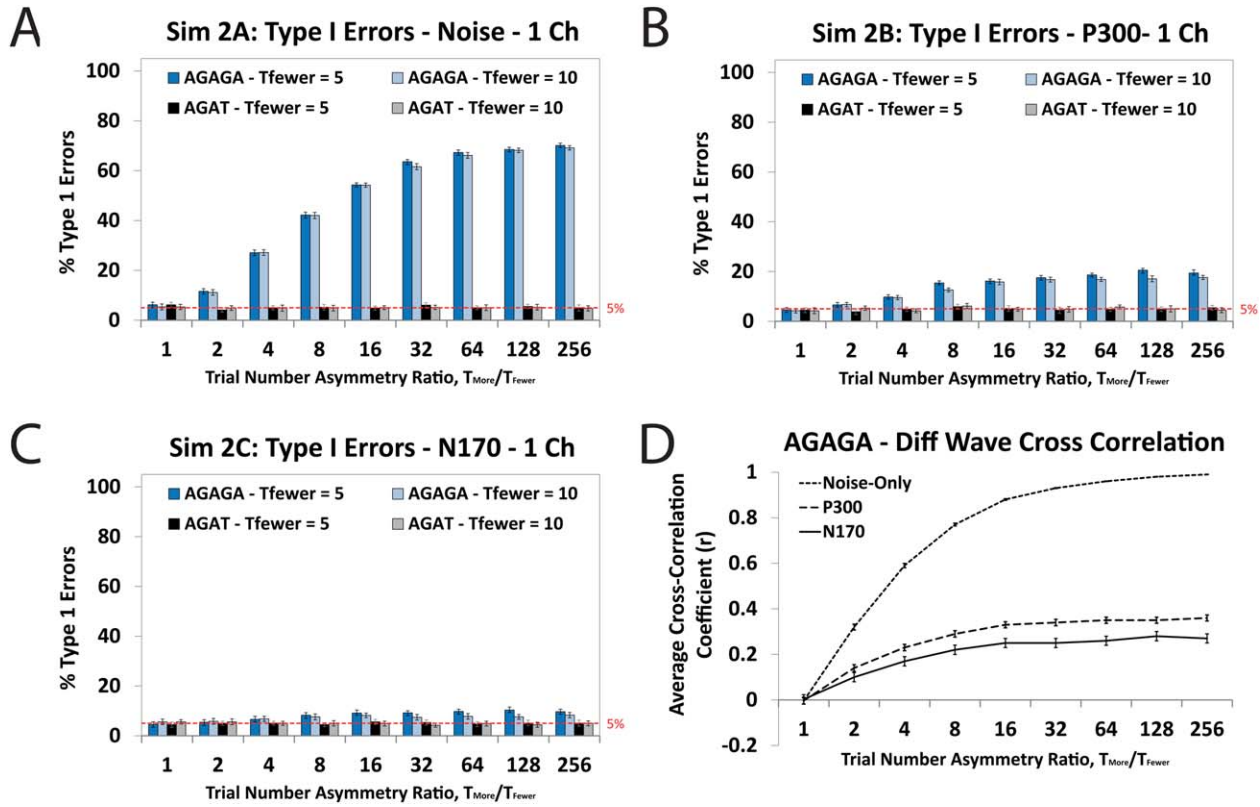

Figure 3. Simulation 2 results. For Panels A–C, the horizontal dashed red line indicates the target 5% Type I error rate level and error bars represent 95% CIs (same method as Simulation 1 methods but with 1,000 replicates). A: Simulation 2A Type I error rates are plotted as a function of trial number asymmetry ratio, T_{More}/T_{Fewer}, when using either the AGAGA (blue bars) or the AGAT (black and gray bars) for ROI selection in noise-only data. Dark blue and black bars represent simulations with $T_{Fewer} = 5$, whereas the light blue and gray bars represent simulations with $T_{Fewer} = 10$. The results show that the AGAT remains unbiased for ROI selection across all condition trial number asymmetries tested, whereas the nonindependent AGAGA becomes increasingly biased as trial asymmetry increases. The results do not depend on the absolute number of trials as different values of T_{Fewer} produce the same results (cf. dark and light bars). The difference wave ROI produced approximately 70% errors regardless of T_{More}/T_{Fewer} level and is not plotted. B: Simulation 2B Type I error rates as a function of trial number asymmetry ratio for data containing noise plus P300 ERP signal. C: Simulation 2C Type I error rates as a function of trial number asymmetry ratio for data containing noise plus N170 ERP signal. D: Average cross-correlation *r* values between the difference wave and the AGAGA for Simulations 2A–C are plotted as a function of condition trial number asymmetry ratio, T_{More}/T_{Fewer}, for noise-only data (2A, dotted line), P300 data (2B, dashed line), and N170 data (2C, solid line). These show higher cross-correlation between AGAGA and difference wave with increasing trial number asymmetry ratio. Error bars represent 95% CIs of each distribution.

Method

We generated data for Simulation 2 with the parameters used in Simulations 1A–C (noise-only; noise+P300, noise+N170) except that we varied the ratio of the number of trials in the two conditions: T_{More} (number of trials in the condition with more trials) and T_{Fewer} (number in the condition with fewer trials). The resulting condition trial number asymmetry was expressed as a condition trial number asymmetry ratio, T_{More}/T_{Fewer}. For computational efficiency, we reduced the base number of trials from 50 per condition (as in Simulation 1) to 10. Thus, for the ratio $T_{More}/T_{Fewer} = 1$, the simulation contained 10 trials per condition. For the other trial asymmetries, T_{Fewer} was always 10 trials whereas T_{More} took values of $T_{Fewer} \times 2^i$ with $i = 0$ to 8. This resulted in condition trial number asymmetry ratios, T_{More}/T_{Fewer}, of 1, 2, 4, 8, 16, 32, 64, 128, and 256 (i.e., T_{More} was 20, 40, 80, 160, 320, 640, 1,280, 2,560 trials, respectively). To test whether the ratio of trial numbers in the two conditions is the determining factor, rather than the absolute number of trials, we also repeated all of these simulations with half the number of trials ($T_{Fewer} = 5$), but with the same trial asymmetry ratio values (see Figure 3A–C, black and dark blue bars). The number of experiments per simulation was reduced to 1,500 for computational efficiency. Thus, for each level of trial number asymmetry within each simulation, there were 1,500 experiments conducted.

As in Simulations 1A–C, we chose ROI positions on the difference wave, the AGAGA, and the AGAT wave, separately. ROIs were chosen as the minimum for Simulation 2A, maximum for 2B (noise-only simulations and P300) and the minimum for Simulation 2C (N170). For all of these ROIs, we calculated the Type I error rate as the percentage of experiments with a significant difference between conditions. Only results for the peak (integration window size = 1) are shown to reduce figure complexity because results for the AGAT were equivalent across integration window sizes.

Cross-correlations between the difference wave and the AGAT and AGAGA, separately, were computed to assess the independence of AGAGA and AGAT from the difference wave. All cross-correlations were assessed at zero-lag. A distribution of cross-correlation *r* values was determined separately for AGAGA and AGAT. The mean *r* value was computed for each simulation, and 95% CIs for the correlations were generated based on the standard deviation and the sample size: $\pm 1.96 * (SD/\sqrt{n})$.

Table 1. *Simulation 2C (N170) Multichannel Type I Error Rates*

	Trial number asymmetry ratio, T_{More}/T_{Fewer}								
Number of channels	1	2	4	8	16	32	64	128	256
AGAGA									
8	5.1%	9.7%	23.1%	36.4%	58.5%	72.1%	91.2%	100%	100%
16	4.8%	12.9%	30.2%	48.4%	68.4%	85.4%	100%	100%	100%
32	4.6%	21.5%	55.7%	85.4%	99.9%	100%	100%	100%	100%
AGAT									
8	5.0%	4.8%	5.3%	4.5%	4.2%	5.1%	4.5%	4.6%	5.0%
16	5.3%	4.9%	5.1%	4.7%	3.8%	4.8%	4.7%	3.9%	3.9%
32	3.8%	4.7%	5.1%	4.5%	4.5%	4.6%	5.1%	4.7%	3.9%

Results and Discussion

As the condition trial number asymmetry ratio, T_{More}/T_{Fewer} (ratio of condition with more trials to condition with fewer trials), increased, the cross-correlation of the AGAGA with the difference wave also increased (Figure 3D). This nonindependence was stronger for noise-only data (Figure 3D, dotted line) than for data containing ERP signals (Figure 3D, dashed and solid lines) presumably because the ERP signals introduced variance that was not different between conditions. As would be expected from using a nonindependent waveform for selection, Type I error rate for AGAGA-based ROIs increased with trial number asymmetry ratio (Figure 3A–C, blue bars) for all three simulations. However, these increases were substantially attenuated by the presence of ERP deflections in the data (Figure 3B,C) compared to pure noise data (Figure 3A). All of the results in Figure 3 represent data with one channel. Results for multichannel data show a similar increase but with higher overall error rates (Table 1, N170, but results were equivalent for P300 and noise-only data).

In contrast, the AGAT was not correlated with the difference wave at any of the trial number asymmetries that we tested (average cross-correlation for all data types, $r = .002$) and regardless of whether the data contained ERP deflections or pure EEG noise. Furthermore, the Type I error rate remained at 5% when using the AGAT for ROI selection (Figure 3A–C, black and gray bars) for all of the condition trial number asymmetry ratios, T_{More}/T_{Fewer}. This was also true for multichannel data N170 data (Table 1), and there were equivalent results for P300 and noise-only data.

Simulations involving different numbers of trials, but having the same trial number asymmetry ratios, showed exactly the same results (Figure 3A–C, compare black and gray bars). This indicates that the trial asymmetry ratio, rather than the total number of trials, drove the bias within the AGAGA results and that the AGAT is robustly safe in the presence of trial number asymmetries regardless of the total number of trials.

The results of Simulation 2 demonstrate that the AGAT is robust to between-conditions trial number asymmetries for all of the asymmetry ratios that we tested. We anticipate that these ratios far exceed those that would be encountered in actual experiments, and thus the AGAT can be treated as essentially unbiased for all practical purposes. It is important to note that the AGAGA was not independent of condition differences when between-condition trial number asymmetries were present.

Simulation 3: Condition Noise Asymmetry

Although AGAT-based ROI selection is robust to condition trial number asymmetries, an asymmetry of noise between conditions could render the AGAT nonindependent (Kilner, 2014) under the null hypothesis (i.e., no mean difference). To systematically test this, we generated simulations with the same parameters as in Simulations 1A–C (including equal trial numbers in the conditions) except that we varied the ratio of the noise in the two conditions (i.e., condition noise asymmetry ratio, $Noise_{Higher}/Noise_{Lower}$).

Method

The parameters for these simulations were exactly the same as those for Simulations 1A–C except that we varied the ratio of the noise amplitude in the two conditions. To generalize our findings beyond the total noise levels, we expressed the noise asymmetry as a ratio of condition noises, N_{Higher} and N_{Lower}, and called this the *condition noise asymmetry ratio*, N_{Higher}/N_{Lower}. In the case of $N_{Higher}/N_{Lower} = 1$ (equal noise), the simulations were replications of Simulations 1A–C. For the other simulations, $Noise_{Higher}$ took values of $Noise_{Lower} \times 2^i$ with $i = 0$ to 11. This resulted in condition noise asymmetry ratios, N_{Higher}/N_{Lower}, of 1, 2, 4, 8, 16, 32, 64, 128, 256, 1,024, and 2,048 (see horizontal axes in Figure 4). ROIs were selected on the difference wave and AGAT as in Simulations 1A–C and additionally on the AGAGA. Simulation 3A contained noise-only, 3B was noise+P300, and 3C was noise+N170. Only results for the peak (integration window size = 1) are shown to reduce figure complexity. As integration window increased, the pattern was similar to the peak results but with lower overall Type I error rates. To reduce computing time, we reduced the number of experiments used to generate each data point to 1,500 rather than the 10,000 used in Simulation 1.

In an additional Simulation 3D (Figure 4D), to examine the effect of ERP signal amplitude on Type I errors for the AGAT/AGAGA, we varied the amplitude of the ERP signal within the data for noise+N170 data only. At 100% amplitude, the N170 negative polarity peak reached −8 and the simulation was equivalent to Simulation 3C. At 0% amplitude, there was no ERP signal present in the data, and the simulation was equivalent to Simulation 3A. The N170 signal was scaled in increments of 20% between these values and the Type I error rate estimated across the different condition noise asymmetries.

Results and Discussion

In Simulation 3A (noise-only), Type I error rates for AGAT-based ROIs increased with condition noise asymmetry (Figure 4A, black line). At asymmetry ratios above approximately $N_{Higher}/N_{Lower} = 8$, error rates for AGAT-based ROIs were equivalent to rates for difference wave ROIs (Figure 4A, gray line). A similar pattern of results was seen for Simulation 3B and 3C (containing P300 and N170 ERP signals, respectively), except that AGAT error rates

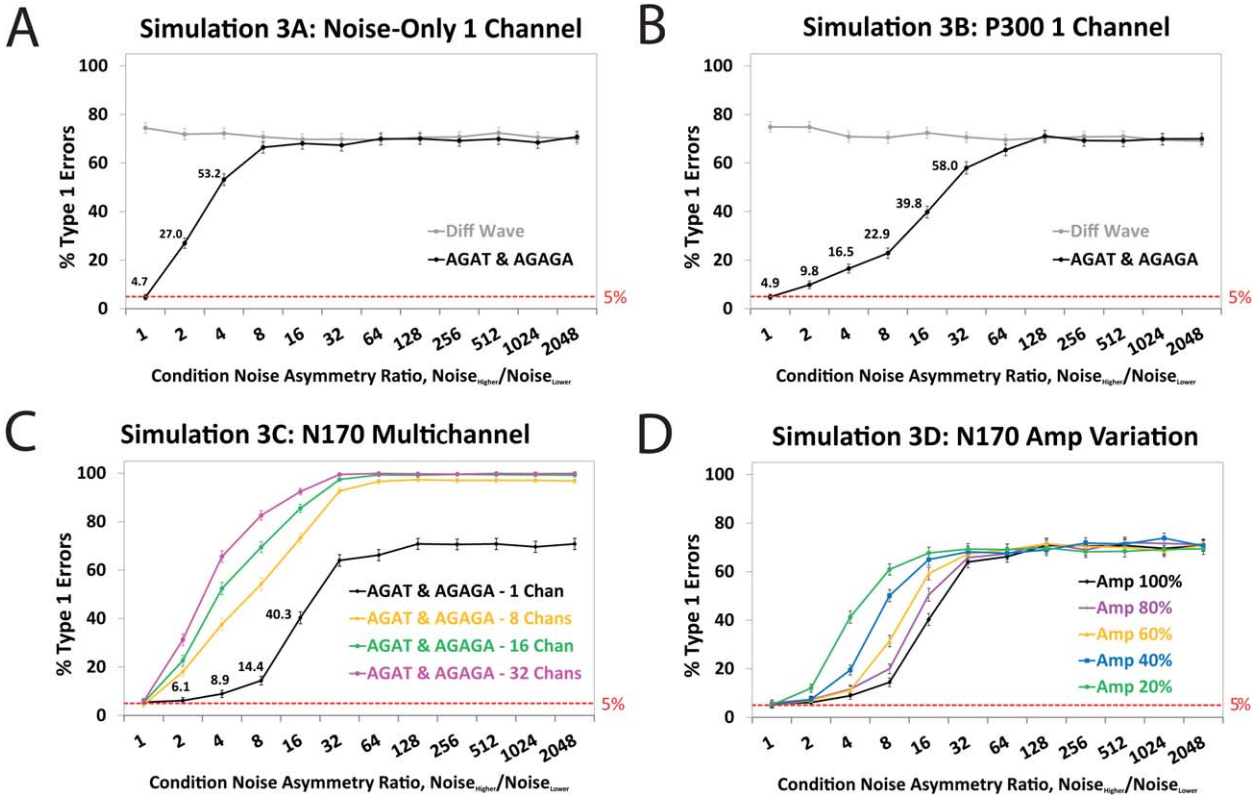

Figure 4. Simulation 3 results. Simulations 3A–D examined the effect of a condition noise amplitude asymmetry on Type I error rates and compared three ROI selection methods. Type I error rate is plotted as a function of the condition noise asymmetry ratio, $Noise_{Higher}/Noise_{Lower}$. Higher values mean a larger asymmetry. Error bars represent 95% CIs (same method as Simulation 1 methods but with 1,000 replicates). A: The results for Simulation 3A (noise-only data) showed that Type I error rates were high for ROIs based on the difference wave (gray line) regardless of noise asymmetry level. ROIs based on the AGAT and AGAGA produced identical results and thus only one line is plotted for these (black line). Type I error rates for AGAT and AGAGA ROIs increased with condition noise asymmetry. B: Simulation results for Simulation 3B, condition noise asymmetry with noise+P300 data. The addition of ERP signal reduced Type I error inflation but bias remained and increased with noise asymmetry. C: Simulation results for Simulation 3C, condition noise asymmetry with noise+N170 single-channel data (black line). Searching for the ROI across time and space in multichannel data further increases the Type I error rates (8 channels, yellow; 16 channels, green; 32 channels, pink). D: In Simulation 3D (single-channel data), the amplitude of the N170 ERP signal was varied from 20% (green line) to 100% (black line), equivalent to panel C (black line) of the 8 μV used in the other simulations in increments of 20% (other colored lines, see legend). Resistance to inflation of Type I error rate increased with increasing amplitude of the ERP signal (i.e., increasing signal-to-noise ratio of the feature of interest).

(Figure 4B,C, black lines) were lower than those for noise-only data and approached the difference wave level (Figure 4A,B, gray lines) at higher asymmetry ratios (above $N_{Higher}/N_{Lower} = 32$). AGAGA and AGAT produced exactly the same results, and thus only one line was plotted for these.

Results for simulations with multiple channels in the N170 simulation show that, in the presence of a condition noise asymmetry, Type I error rates increased for AGAT-based ROIs as the number of channels increased (Figure 4C, colored lines vs. 1-channel black line). The impact of multiple channels was similar for N170, P300 and noise-only data. Thus, to reduce figure complexity, we plotted multichannel data only for the N170.

The results of Simulation 3 place an important constraint on the use of the AGAT. In cases of asymmetric condition noise, the AGAT can be biased to the exact same extent as the AGAGA. This is different than with condition trial number asymmetries (Simulation 2) where only the AGAGA was biased. The amount of bias depends on the signal-to-noise ratio of the ERP feature of interest (i.e., N170 or P300 peak, in this case). This was shown in a series of further simulations (with N170 ERP). As the ERP peak amplitude was increased from 0 to the full intensity (8 μV max), the absolute levels of bias decreased (Figure 4D). Thus higher signal-to-noise ratio ERP peaks were more shielded, though not completely, from the bias than lower signal-to-noise ratio peaks. Although the peak amplitudes were the same for the N170 and the P300, it is clear that there were some small differences in susceptibility to bias across the condition noise asymmetry range (cf. shape of black lines, Figure 4B,C). These could signal that the AGAT's bias depends slightly on the type of ERP peak or feature of interest even when they have the same signal-to-noise ratio. However, further work will need to be done to determine exactly which factors affect this. Finally, the absolute level of bias increased with the number of channels in the data across which the search for the AGAT peak was conducted (Figure 4C).

In Simulation 3, we found that the average zero-lag cross-correlation between the AGAT and the difference wave increased as a

Table 2. *Simulation 3C (N170) Average Cross-Correlation Between AGAT and Difference Wave for Amplitude = 100%*

Noise asymmetry	1	2	4	8	16	32	64	128	256	512	1,024	2,048
r value	.006	.151	.392	.679	.886	.968	.993	.996	.997	.998	.998	.998

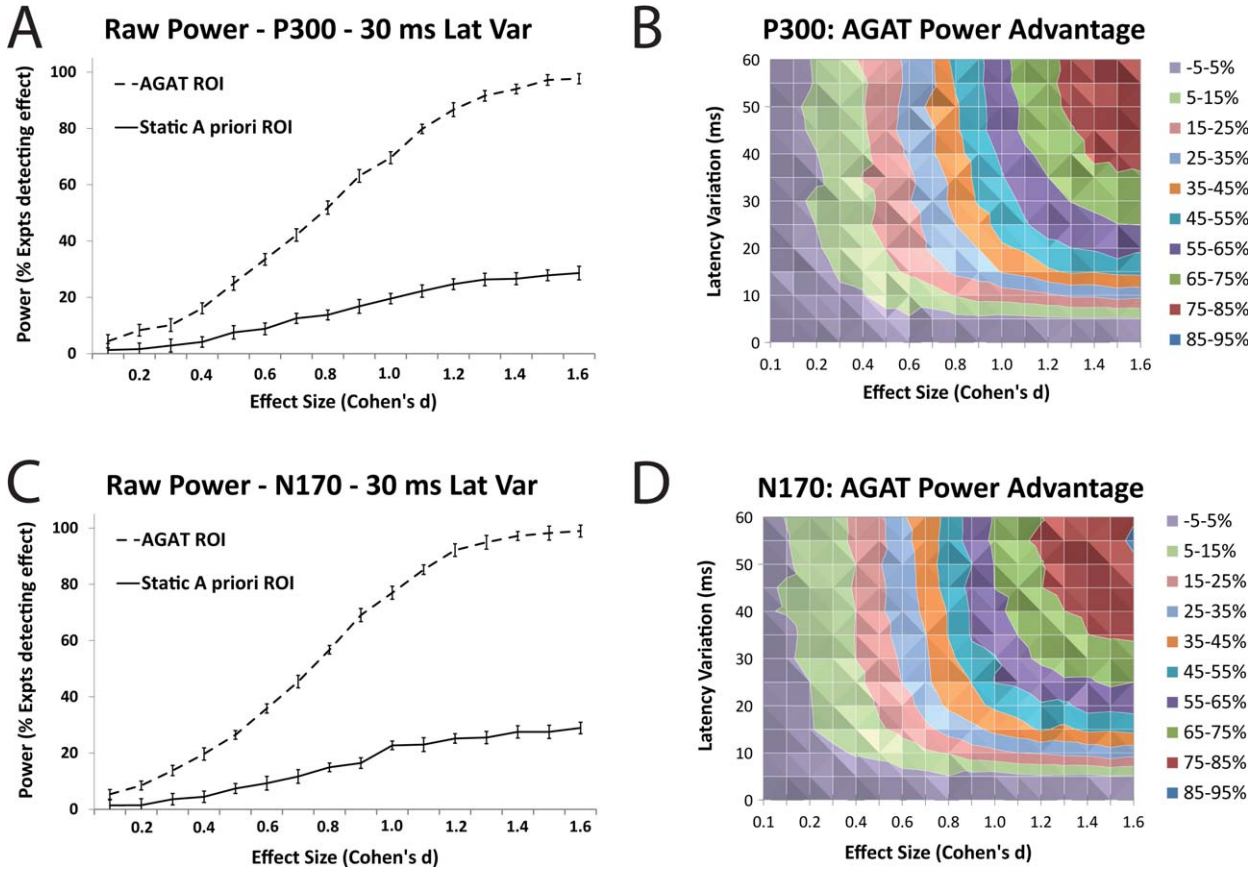

Figure 5. Simulation 4: Raw power and AGAT power advantage. Error bars represent 95% CIs (same method as Simulation 1 methods but with 1,000 replicates). A: Raw power is plotted as a function of effect size (Cohen's *d*) for detecting effects located at a P300 peak using either an AGAT-based ROI selection (dashed line) or an ROI positioned at a static a priori position (solid line). AGAT-based ROIs outperformed a priori ROIs. Power increased with effect size but the increase was larger for AGAT-based ROIs than for a priori ROIs. This is for the simulation in which the latency of the effect varied (across experiments) with a *SD* of 30 ms. B: For simulations with P300 ERP signals, the power advantage of using an AGAT-based ROI (calculated as AGAT-ROI power minus a priori ROI power) is plotted as a function of effect size (Cohen's *d*) and latency variation of the effect (*SD* of latency in ms). Color represents the power advantage (%) as indicated in the legend (e.g., light purple = −5–5% advantage for AGAT). Higher positive values indicate a greater advantage of AGAT. The advantage of AGAT-based ROIs increased with both latency variation and effect size. This plot includes the data from (A), which is a horizontal slice at the latency = 30-ms level representing the difference between the lines plotted in (A). C: Raw power is plotted as in (A) but for data containing an N170 ERP signal with the effect located near the N170 peak. The results are the same as for P300 data. D: AGAT power advantage is plotted as in (B) but for data containing an N170 ERP signal and show the same advantage of using AGAT-based ROIs as for P300 data.

function of the noise asymmetry (Table 2) in a manner similar to that seen for trial number asymmetry in Simulation 2. Eventually, this correlation approached $r = 1$ at higher asymmetry values. This means that the AGAT, at high noise asymmetry, comes to almost perfectly match the difference wave. Using an AGAT that closely reflects the differences wave inflates Type I error rate substantially because the difference wave is not independent of the contrast of interest.

Overall, the results of Simulation 3 suggest that the AGAT is not safe to use when the amplitude of the individual trial EEG noise differs between conditions. Even at our lowest noise asymmetry of 2 (double noise in one condition compared to the other), we could find Type I error rates of up to 30% when selecting the AGAT among multiple channels in a high amplitude component (32-channel N170 data). Although some protection against Type I errors seems to be afforded by using high signal-to-noise ratio ERP features/peaks, further work is needed to determine the full range of parameters that need to be considered. We advise against using the AGAT when condition noise asymmetry is greater than 1.5, especially in multichannel data or when considering ERP features with lower signal-to-noise ratios than used in our simulations (approximate signal-to-noise ratio = 0.4 in our 100% case, see methods for noise and signal amplitudes).

Simulation 4: AGAT Power

It is clear from Simulations 1–3 that AGAT-based ROI selection can avoid inflating Type I error rate. However, does using the AGAT to position ROIs actually adapt to the features of the data, and thus potentially increase power, as we suggested above? In order to evaluate this, we conducted power simulations and compared AGAT-based ROI selection with the commonly used method of selecting an ROI based on a priori or independent information. We hypothesized that using the AGAT would be advantageous because, assuming that the location of effects varies between experiments, the AGAT, being data driven, should take account of experiment-specific data features whereas a priori/independent information cannot.

To assess power, we generated data as in Simulations 1B and 1C (noise+ERP) but with two differences. First, we varied the latency of the ERP (P300 and N170) peaks across experiments within each simulation to simulate experiment-to-experiment variation of ERP peak latencies. If this variation is large, then we expected a priori/independent ROIs to regularly miss effects because they cannot take this variation into consideration. In contrast, the AGAT should detect the relevant peak in each experiment regardless of the variation across experiments, giving it an advantage at higher levels of variation.

Second, at the relevant peak (N170 or P300), we inserted a difference between conditions. The size of this effect varied across simulations. In each experiment within a simulation, we then conducted hypothesis tests at two ROIs. One ROI was an a priori/independent ROI that was the same for all experiments within a simulation (i.e., the middle of the latency distribution for the ERP peak of interest). The other ROI was selected by using the AGAT to find the N170 or P300 peak. We then estimated the power, that is, the percentage of correctly detected effects for each ROI. For simplicity, Simulation 4 was conducted with a single channel of data.

Table 3. *Steps for Selecting an AGAT-Based ROI Position*

Step	Instructions
Step 1	Aggregate all trials from all conditions and all participants into one set. Do not use subject ERPs or condition grand averages.
Step 2	Average waveforms/maps across this set of trials to generate the aggregate grand average from trials (AGAT) waveform.
Step 3	Select a peak (or other feature) of interest on this waveform (e.g., for the N170 this may be a minimum between 150–200 ms). This must be selected a priori and should not be changed based on statistical testing of the difference between conditions.
Step 4	Apply your integration window, or other quantification method, of choice (based on a priori information) and perform statistical analysis, as usual, at this location on original data.

Method

Data were generated as in the single-channel versions of Simulations 1B and 1C (noise+P300, noise+N170, respectively) except that we varied two things. First, at the ERP peak location (maximum for P300, 200 ms; minimum for N170, 477 ms), we added a boxcar effect (difference between conditions) lasting 21 samples (21 ms) and centered on the peak. This was added to one condition. The other condition was unchanged relative to Simulation 1. Due to the different peak polarities for the two ERP components, for the P300 simulation (Simulation 4A), positive effect values were added; whereas for the N170 simulation (Simulation 4B), a negative effect was added. This simulated an amplitude increase of the peak in one condition compared to the other. Although not realistic, a boxcar effect allowed us to have uniform effect size across the effect interval. This was important in giving validity to our manipulation of effect size across simulations. Otherwise, effect size would have varied across time within each experiment within the simulation.

Across simulations, we varied the amplitude of this effect across 16 levels: 0.03125, 0.06250, 0.09375, 0.12500, 0.15625, 0.18750, 0.21875, 0.25000, 0.28125, 0.31250, 0.34375, 0.37500, 0.40625, 0.43750, 0.46875, and 0.50000 μV. These effect amplitudes were chosen to correspond to a particular set of effect sizes (Cohen's *d*) ranging from 0.1–1.6 in increments of 0.1. For each effect amplitude, we calculated the corresponding effect size (Cohen's *d*) by dividing the effect amplitude by the average within-condition noise. The within-condition noise was estimated from the simulated data. Within one condition (without added effect) of each simulated experiment, we calculated the standard deviation of the participant ERP amplitudes at the selected ROI (peak only, one sample window). The average within-condition noise across all experiments was approximately 2.5 μV. Effect size values are used as the *x* axes in Figure 5to provide generality of the results across experiments with different absolute levels of noise and effect amplitudes.

The second change from Simulations 1B, C involved addition of latency variation of the ERP peaks. This was achieved by shifting the entire ERP waveform left or right and padding with zeros. Latency varied according to a normal distribution centered on the original peak location (N170 = 200 ms; P300 = 477 ms). Across simulations, we varied the standard deviation of the latencies from 0 (no variation, as in Simulations 1–3) to 60 ms (in 5-ms steps). Thus, we conducted 208 simulations (16 Effect Sizes × 13 Latency *SD*s) each for the two ERP components. To reduce total processing time, each simulation included 1,500 experiments (instead of 10,000 in Simulation 1). For each experiment within a simulation, we conducted a hypothesis test at each ROI and then counted the percentage of experiments in which an effect was significantly detected within the time range of the inserted effect (i.e., power).

Results and Discussion

Figure 5A,C show the raw power for AGAT (dashed line) and a priori (solid line) ROIs as a function of effect size when the average latency variation of the peak was 30 ms. The AGAT consistently had higher power than the a priori ROI, especially at higher effect sizes. Because we were primarily interested in the difference in power between AGAT and a priori ROIs, we calculated the difference in power between them (AGAT minus a priori) for each simulation and plotted this difference, the AGAT power advantage, as a function of effect size and latency variation (Figure 5B,D). Higher positive values indicate that AGAT had greater power than a priori ROIs, and negative values would indicate the reverse. Values of zero indicate equivalent power. In data with low latency variation (below 5–10 ms, on average), AGAT and a priori methods had approximately equal power (Figure 5B,D). However, when latency variation was 15 ms or greater, the AGAT became substantially more powerful than a priori methods at effect amplitudes above 0.3 (Figure 5B,D). It is important to note that this simulation was carried out, for simplicity of design, with single-channel data. Thus, strictly speaking, we cannot generalize the exact size of the AGAT benefit to situations when one may also be identifying an ROI position on a multichannel AGAT. However, we expect that the benefit of AGAT over the independent ROI will hold across multichannel data because the AGAT should allow adaptation to changes in the location of the peak in space/channel in addition to changes in latency (as we have shown in Simulation 4). This is because the feature of interest (peak here) can be detected across space as well as in time. In contrast, an a priori/independent ROI cannot, by definition, show this adaptability and thus should have less power to detect the effect. However, further work will be required to confirm and quantify this benefit.

Table 4. *AGAT Usage Guidelines and Assumptions*

Assumptions/criteria to check	Detail	For more detail
Noise equivalence	The single-trial EEG noise must be approximately equivalent across your conditions. As a rule of thumb, if the noise amplitude is more than 1.5 times greater in one condition than others, then avoid using the AGAT. Note that having unequal numbers of trials in the two conditions does not create this problem (see Simulation 2).	Simulation 3 & Figure 4
AGAT method of computation	The AGAT must be computed from the individual trials of all participants and not from the participant ERPs.	Simulation 1 methods
Latency equivalence	The latency of your ERP feature of interest (usually a peak) must be approximately equivalent across your conditions. If you expect or see significant latency differences, AGAT may not be appropriate	General discussion, paragraph 2
Waveform morphology equivalence	The morphology of the ERP waveform must be approximately equivalent across conditions. A failure of this assumption could reduce power or produce misleading results.	General discussion, paragraph 2
ERP feature of interest is known	You must have an a priori hypothesis about which ERP feature you intend to locate and have a priori criteria for detecting it on the AGAT. For instance, this may be a particular peak and you must specify the polarity and other criteria (e.g., negative polarity peak/minimum between 150 and 220 ms.) If little or no information is known, then mass univariate methods may be more appropriate.	Simulation 1 methods & general discussion, paragraph 2
Expected latency variation	The AGAT confers the biggest advantage over a priori/independent ROI selection when the variation in latency of the ERP feature across experiments is higher. Features with less latency variability benefit less.	Simulation 4 & Figure 5

General Discussion

We have demonstrated empirically that ROIs can be selected in a data-driven manner without inflating Type I error rates by selecting peaks of the AGAT. This method is safe even in the presence of an asymmetry in the number of trials between the conditions.[2] However, this is subject to two conditions. First, the AGAT must be computed by averaging the aggregate of all individual trials from both conditions rather than averaging over grand averages (AGAGA). Secondly, using the AGAT with large condition noise asymmetries can inflate Type I error rates. This could occur, for instance, when comparing data from a patient group with control participants. Our results show that, even with relatively small noise asymmetries (e.g., $\times$ 2), Type I error rates can inflate to 6.1% (N170, Figure 4C) and (9.8%, Figure 4B) in single-channel data and further in multichannel data. It is clear that higher signal-to-noise ratio/amplitude of the ERP peak of interest can partially protect against this at low noise asymmetries (Figure 4D). However, a more detailed exploration of this will be needed to identify all of the relevant factors. Finally, our power simulations showed that, subject to certain assumptions (see following), using the AGAT for ROI selection can be more powerful than a common method of selecting ROIs based on a priori/independent information. Thus, we believe that using the data-driven AGAT for ROI selection is a safe and effective method when one is looking for ERP features, such as peaks, at which to position an ROI for testing. It allows one to take advantage of more information in the data to customize ROIs to its features. Table 3 provides an outline of the steps that should be used to calculate the AGAT for use in studies.

The AGAT is not appropriate for all data and analyses. Our results have already highlighted that differences in noise amplitude between the conditions can introduce bias. Additionally, using the AGAT depends on two key assumptions: (1) the effect of interest will have approximately the same latency across all of the aggregated conditions, and (2) the morphology of the ERP waveform is approximately the same across all conditions. If this is not the case, then the power of the AGAT will likely be significantly reduced or the results could be misleading. This arises because when there are significant latency or ERP morphology differences between conditions, then aggregating across them may create an AGAT waveform with peaks or other features that are not present in all, or any, of the individual conditions. Thus, the ROI would miss the effect. However, it is worth pointing out that this assumption applies equally to ROI selection based on a priori/independent information unless it explicitly takes into account latency/morphology differences between conditions. Finally, the AGAT will be of no use in analysis if there is no a priori hypothesis about which peak/feature of the AGAT is relevant. The researcher must provide a rule for choosing the peak, or other feature, on the AGAT. In cases where there is no or little information about the location of effects,

2. The following is one observation about why the AGAT is unbiased under trial number asymmetry. Assume, X trials for Condition A and Y trials for Condition B, with $X > Y$. The peak (or peak interval) selected in the AGAT is (in a statistical sense) biased more toward Condition A's actual peak than Condition B's. However, this disparity in bias is counteracted by the disparity in ERP amplitude due to averaging (i.e., amplitudes in Condition A ERP are, in a statistical sense, lower, or less extreme, than in Condition B, since A involves averaging more trials).

researchers may want to consider mass univariate (Blair & Karniski, 1993; Groppe, Urbach, & Kutas, 2011; Kilner, Kiebel, & Friston, 2005; Maris & Oostenveld, 2007) and multivariate (Hemmelmann et al., 2004; McIntosh & Lobaugh, 2004) approaches where one can analyze across large portions of a data set (with appropriate correction). Although the ability of mass univariate approaches to detect unexpected effects while controlling Type I errors is an incredibly useful complementary tool to ROI-based analysis, many of these methods require substantial experience, specification of a number of parameters for analysis, and some cost to power. Furthermore, we expect that, when an effect is typically known to occur near a localizable AGAT data feature (e.g., peak) and it is of low to medium effect size, AGAT-based ROI methods will be more powerful than mass univariate methods. However, a more detailed comparison between the power of AGAT and mass univariate methods will require further work across the range of different mass univariate methods to confirm this. When there is a clear prediction about which peak/feature along the AGAT will be associated with the effect, we believe that AGAT-based ROI approach should be preferred. Table 4 provides a summary of the factors that researchers should check to determine whether using the AGAT is likely to be safe and powerful for their data.

Assuming that ERP features of interest (peaks here) vary from one experiment to the next, as we simulated, and that the effect is colocated with that feature, our results suggest that using AGAT-based ROIs can be more powerful than a priori ROIs. This is because, unlike an a priori ROI, the AGAT contains experiment-specific information about the latency of ERP features and can be used to position tests at that location. Importantly, in our results, the AGAT never performed worse than the a priori method. Use of the AGAT does assume that the effect of interest is colocated with a feature of interest on the AGAT waveform. If this is not the case, then use of the AGAT will not be an effective way of localizing the ROI. However, we believe that, in many cases, researchers already assume that this is the case and do aim to position ROIs at a particular peak or other feature.

Other researchers have previously suggested something like the AGAT for ROI selection in other domains (Keil et al., 2014; Kilner, 2013, 2014; Kriegeskorte et al., 2009; Luck, 2014), and our informal discussions with ERP researchers suggest that some already use data-driven methods such as the AGAGA. In reviewing the method sections of 20 randomly selected N170 ERP papers, it is clear that some researchers localize peaks on grand-averaged data for quantification. However, it is often not clear from the reported methods how they aggregated their data (i.e., AGAT, AGAGA, or otherwise) and whether independence was established. We hope that our results and further discussion of this issue will prompt researchers to more clearly report their ROI-selection procedures and reviewers to request this information.

In our work, we have focused on identifying peaks on the AGAT because these are ERP features that, in our reading of the literature, are commonly used for analysis, and they are easily identified. However, as others have pointed out (Luck, 2005), voltage peaks in the ERP waveform are not equivalent to ERP components and do not necessarily reflect the underlying latent ERP components in which researchers are interested. We acknowledge this and encourage researchers to consider alternative methods of quantification (Luck, 2014). However, our goal is not to provide an analysis of these issues here. Given that researchers can and do commonly use peaks to localize and quantify ERP components, our goal was to analyze how to do this with high power and without inflating Type I errors. Furthermore, we believe that, in principle, other features (e.g., largest area under the curve, zero crossings) of the AGAT may be valid for unbiased ROI localization. Additionally, in our work, we have always selected the absolute maximum and minimum peaks across the waveform. However, we see no reason, in principle, why selecting a lower amplitude, local (within a search window) peak within the AGAT waveform, which may be more appropriate for other ERP components (e.g., P1, P2), should be any different as long as the AGAT is used for selection and the assumptions of use are met (see second paragraph of General Discussion above and Table 4). However, this will need to be confirmed with further work. In particular, the power of AGAT when selecting nonpeak or lower amplitude features will need to be assessed in greater detail and compared to ROIs based on independent data and other methods.

Although we have focused on using the AGAT in ERP studies, this approach can be applied more widely. In principle, one can also use AGAT-based ROI selection in EEG/MEG time-frequency studies, eye tracking fixation probability maps (Caldara & Miellet, 2011), psychophysiological measures, and other types of multidimensional data. There is no reason, in principle, to believe that adding further dimensions to the data should render the AGAT biased. In fact, fMRI researchers often use orthogonal comparisons in 3D data sets (or independent data) to generate ROIs for analysis, and there has been substantial discussion of this practice (Friston, Rotshtein, Geng, Sterzer, & Henson, 2006; Kriegeskorte et al., 2009; Nieto-Castañón & Fedorenko, 2012; Poldrack, 2007; Saxe, Brett, & Kanwisher, 2006; Vul et al., 2009). In our analysis, we selected ROIs in the time dimension but the AGAT can also be computed across spatiotemporal ERP data as well.

In practical terms, nearly all ERP analysis software should allow calculation of the AGAT. However, this may depart significantly from the typical ERP processing pipeline and be cumbersome in some software. One barrier will be that ERP analysis software does not typically involve averaging individual trials across participants. This is because it is common first to compute the ERP average for each participant separately and, only then, compute the grand average of participants' ERPs (i.e., steps toward computing the AGAGA but not the AGAT). For instance, MATLAB-based FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011) and ERPLAB (Lopez-Calderon & Luck, 2014), to our knowledge, do not automatically allow segments from different participants to be averaged together without first creating an ERP[3] (a step which is prohibited in calculation of the AGAT). BrainVision Analyzer (Brain Products, GmbH; http://www.brainproducts.com/) does allow computation of the AGAT, but only using the weighted average option within its grand average function.[4] For other software, researchers should check carefully exactly how their existing averaging functions work to determine whether they support the AGAT. With some programming skill, it is possible to add one's own functions to these packages to overcome this. However, one simple and immediately available way around this constraint in all three software packages above is

3. The compute average ERPs (pop_averager) function in ERPLAB allows more than one data set/participant to be selected when computing an ERP. However, based on a personal communication (April 2016) with the ERPLAB developers, this function first computes the ERP for each participant and then computes the grand average of these ERPs. Thus, it does not meet the requirements for computing the AGAT.

4. Based on a personal communication with Brain Products technical support (support@brainproducts.com), using the grand average function with the "calculate weighted average" box ticked will compute the AGAT as a weighted average of all of the individual trials from participants.

to append all of the participant data files together into one long file (e.g., *ft_appenddata* function in FieldTrip; Append File option in BrainVision Analyzer) and then do segmentation (combining data from all trial types into one condition label) and averaging across segments/epochs within this multiparticipant file (which contains all participants' individual trials). Once the AGAT waveform has been computed, the time/location of the feature of interest (a peak in our examples) can then be found either by visual inspection of the AGAT (with clear a priori criteria) or by using, for instance, a peak detection function (with appropriate polarity and approximate time/location criteria). The result can then be used as the exact position of the ROI, and quantification of the data can go forward as with any other ROI analysis in the original data set.

Although some data-driven methods for data analysis have been shown to be biased, not all are problematic. Our results demonstrate a simple, unbiased, data-driven method for ROI localization for ERP data that can likely be generalized more broadly. Using data-driven methods such as the AGAT may also increase power to detect effects when effect latencies vary from experiment to experiment avoiding Type II errors. In avoiding Type I errors associated with some data-driven ROI techniques, researchers may be ignoring useful information in data and unnecessarily inflating Type II errors. Most importantly, our results expand our understanding of the conditions under which this particular method of ROI localization can fail and indicate how it needs to be computed in order to minimize bias.

References

Blair, C. R., & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, *30*, 518–524. doi: 10.1111/j.1469-8986.1993.tb02075.x

Bowman, H., Filetti, M., Janssen, D., Su, L., Alsufyani, A., & Wyble, B. (2013). Subliminal salience search illustrated: EEG identity and deception detection on the fringe of awareness. *PLOS ONE*, *8*(1), e54258. doi: 10.1371/journal.pone.0054258

Brisson, B., Robitaille, N., & Jolicoeur, P. (2007). Stimulus intensity affects the latency but not the amplitude of the N2pc. *NeuroReport*, *18*(15), 1627–1630. doi: 10.1097/WNR.0b013e3282f0b559

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi: 10.1038/nrn3475

Caharel, S., Leleu, A., Bernard, C., Viggiano, M.-P., Lalonde, R., & Rebaï, M. (2013). Early holistic face-like processing of Arcimboldo paintings in the right occipito-temporal cortex: Evidence from the N170 ERP component. *International Journal of Psychophysiology*, *90*(2), 157–164. doi: 10.1016/j.ijpsycho.2013.06.024

Caldara, R., & Miellet, S. (2011). iMap: A novel method for statistical fixation mapping of eye movement data. *Behavior Research Methods*, *43*(3), 864–878. doi: 10.3758/s13428-011-0092-x

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science*, *1*(3), 140216–140216. doi: 10.1098/rsos.140216

de Gelder, B., & Stekelenburg, J. J. (2005). Naso-temporal asymmetry of the N170 for processing faces in normal viewers but not in developmental prosopagnosia. *Neuroscience Letters*, *376*(1), 40–45. doi: 10.1016/j.neulet.2004.11.047

Easterbrook, P., Gopalan, R., Berlin, J., & Matthews, D. (1991). Publication bias in clinical research. *Lancet*, *337*(8746), 867–872. doi: 10.1016/0140-6736(91)90201-Y

Flevaris, A. V, Robertson, L. C., & Bentin, S. (2008). Using spatial frequency scales for processing face features and face configuration: An ERP analysis. *Brain Research*, *1194*, 100–109. doi: 10.1016/j.brainres.2007.11.071

Friston, K. J., Rotshtein, P., Geng, J. J., Sterzer, P., & Henson, R. N. (2006). A critique of functional localisers. *NeuroImage*, *30*(4), 1077–1087. doi: 10.1016/j.neuroimage.2005.08.012

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. doi:10.1111/j.1469-8986.2011.01273.x

Hemmelmann, C., Horn, M., Reiterer, S., Schack, B., Süsse, T., & Weiss, S. (2004). Multivariate tests for the evaluation of high-dimensional EEG data. *Journal of Neuroscience Methods*, *139*(1), 111–120. doi: 10.1016/j.jneumeth.2004.04.013

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., … Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, *51*(1), 1–21. doi: 10.1111/psyp.12147

Kilner, J. M. (2013). Bias in a common EEG and MEG statistical analysis and how to avoid it. *Clinical Neurophysiology*, *124*(10), 2062–2063. doi: 10.1016/j.clinph.2013.03.024

Kilner, J. M. (2014). *A note of caution when selecting ROIs from orthogonal contrasts*. Retrieved from http://kilnerlab.blogspot.co.uk/2014/02/a-note-of-caution-when-selecting-rois.html

Kilner, J. M., Kiebel, S. J., & Friston, K. J. (2005). Applications of random field theory to electrophysiology. *Neuroscience Letters*, *374*(3), 174–178. doi: 10.1016/j.neulet.2004.10.052

Koenig, T., Stein, M., Grieder, M., & Kottlow, M. (2014). A tutorial on data-driven methods for statistically assessing ERP topographies. *Brain Topography*, *27*(1), 72–83. doi: 10.1007/s10548-013-0310-1

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. doi: 10.1038/nn.2303

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213. doi: 10.3389/fnhum.2014.00213

Luck, S. J. (2005). Ten simple rules for designing ERP experiments. In T. C. Handy (Ed.), *Event-related potentials: A methods handbook* (pp. 17–32). Cambridge, MA: MIT Press.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd Ed.). Cambridge, MA: MIT Press.

Luck, S. J., & Hillyard, S. A. (1994). Spatial filtering during visual search: Evidence from human electrophysiology. *Journal of Experimental Psychology. Human Perception and Performance*, *20*(5), 1000–1014. doi: 10.1037/0096-1523.20.5.1000

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi: 10.1016/j.jneumeth.2007.03.024

McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances [Supplement 1]. *NeuroImage*, *23*, S250–S263. doi: 10.1016/j.neuroimage.2004.07.020

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*(4), 313–329. doi: 10.1016/0001-6918(78)90006-9

Nieto-Castañón, A., & Fedorenko, E. (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage*, *63*(3), 1646–1669. doi: 10.1016/j.neuroimage.2012.06.065

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 156869. doi: 10.1155/2011/156869

Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, *2*(1), 67–70. doi: 10.1093/scan/nsm006

R Development Core Team. (2014). A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. doi: 10.1037/0033-2909.86.3.638

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, *30*(4), 1088–1096; discussion 1097–1099. doi: 10.1016/j.neuroimage.2005.12.062

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis

allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632

Ten Caat, M., Lorist, M. M., Bezdan, E., Roerdink, J. B. T. M., & Maurits, N. M. (2008). High-density EEG coherence analysis using functional units applied to mental fatigue. *Journal of Neuroscience Methods*, *171*(2), 271–278. doi: 10.1016/j.jneumeth.2008.03.022

Towler, J., & Eimer, M. (2014). Early stages of perceptual face processing are confined to the contralateral hemisphere: Evidence from the N170 component. *Cortex*, *64C*, 89–101. doi: 10.1016/j.cortex.2014.09.013

von der Malsburg, T., & Angele, B. (2015). The rules of statistics make no exception for reading research: False positive rates in eyetracking studies of reading behavior. In A. Gatt & H. Mitterer (Eds.), *Proceedings of the 21th Architectures and Mechanisms for Language Processing Conference (AMLaP)*. Valetta, Malta.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. doi: 10.1111/j.1745-6924.2009.01125.x

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. doi: 10.1037/a0022790

Yeung, N., Bogacz, R., Holroyd, C. B., & Cohen, J. D. (2004). Detection of synchronized oscillations in the electroencephalogram: An evaluation of methods. *Psychophysiology*, *41*(6), 822–832. doi: 10.1111/j.1469-8986.2004.00239.x

Zhang, W., & Luck, S. J. (2009). Feature-based attention modulates feedforward visual processing. *Nature Neuroscience*, *12*(1), 24–25. doi: 10.1038/nn.2223

(Received January 8, 2016; Accepted May 4, 2016)